A novel method for determining linkage between DNA sequences: hybridization to paired probe arrays

Erik Gentalen* and Mark Chee+

Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

Received November 6, 1998; Revised and Accepted January 26, 1999

ABSTRACT

Cooperative hybridization has been used to establish physical linkage between two loci on a DNA strand. Linkage was detected by hybridization to a new type of high-density oligonucleotide array. Each synthesis location on the array contains a mixture of two different probe sequences. Each of the two probes can hybridize independently to a different target sequence, but if the two target sequences are physically linked there is a cooperative increase in hybridization yield. The ability to create and control non-linear effects raises a host of possibilities for applications of oligonucleotide array hybridization. The method has been used to assign linkage in 50:50 mixtures of DNA containing single nucleotide polymorphisms (SNPs) separated by 17, 693, 1350 and 2038 bp and to reconstruct haplotypes. Other potential uses include increasing the specificity of hybridization in mutation detection and gene expression monitoring applications, determining SNP haplotypes, characterizing repetitive sequences, such as short tandem repeats, and aiding contig assembly in sequencing by hybridization.

INTRODUCTION

Grouping and ordering genomic DNA markers is a useful practice in molecular genetics. It is used to make physical and genetic maps and to establish the relationship between genotype and phenotype. A key principle is that markers in close proximity on the same DNA molecule tend to be inherited together, since the likelihood of a recombination event between two markers decreases as their separation decreases. The quantitative measurement of this phenomenon, linkage disequilibrium and its application to genetic analysis, was pioneered by Sturtevant in 1913 (1). Since then, a wide variety of genetic and molecular mapping techniques have been developed and used to build genomic maps based on both genetic and physical DNA markers. Recombination-based mapping has been supplemented by powerful molecular techniques, such as radiation hybrid mapping (2) and methods that are based on single chromosome copy analysis (3,4) or allele-specific detection (5,6). Other methods for establishing linkage include direct visualization of markers, using, for example, electron microscopy and fluorescence in situ hybridization (FISH) (7,8).

In this paper we describe a conceptually different approach for detecting physical linkage between two loci on a DNA molecule. The approach is based on cooperative hybridization to pairs of oligonucleotide probes (the word ‘probe’ is used throughout this text to refer to the oligonucleotides immobilized on the array surface, which query the solution-phase ‘target’ nucleic acids) arrayed on a 2-dimensional glass surface. Physically linked target sequences hybridize cooperatively to both probes in the pair, whereas unlinked sequences hybridize non-cooperatively. We demonstrate linkage detection with single base mismatch discrimination over distances ranging from a small number of base pairs to ~2 kb. We further show that the approach is generally applicable to a wide variety of sequences.

MATERIALS AND METHODS

Oligonucleotide arrays

DNA arrays were synthesized using 5′-MeNPOC-protected phosphoramidites (9,10). A MeNPOC-protected hexaethyleneglycol phosphoramidite is coupled to a glass substrate that has been silanated with bis(hydroxyethyl)aminopropytriethoxysilane. The substrate is then exposed to 365 nm light through a lithographic mask. The MeNPOC protecting group is removed by light, so that only the exposed sites become available for coupling. Repeated cycles of light-directed deprotection and nucleoside base coupling allow the efficient parallel synthesis of any desired combination of large numbers of different oligonucleotides.

Paired probe array synthesis

Paired arrays are synthesized in a similar fashion (Fig. 1). The primary difference is that prior to coupling the first nucleoside the array surface is exposed for one half-life of the MeNPOC protecting group, resulting in an approximate 1:1 mixture of protected and deprotected sites. The deprotected sites are coupled with a 5′-DMT-protected nucleoside. The DMT group is stable through the photodeprotection and synthesis cycles. The remaining MeNPOC-protected linker is patterned as described above into a probe array (array of probe 1). Upon completion, these probes are capped (either with 1:1 acetic anhydride:N-methylimidazole in lutidine/THF or by coupling of N,N-diethyl-N,N-diisopropyl phosphoramidite) to prevent further elongation. After the first
were amplified from samples previously sequenced on an ABI prepared using single-stranded asymmetric PCR. These DNAs human mitochondrial DNAs of 160 bases and 2.5 kb were variants listed above. In addition to the oligonucleotide targets, sequences were derived from the position 10 and 27 double are based on 5′N2 were synthesized (Fig. 3). Each site contains a different combination of N1 and N2 for the substituted by A, C, G or T. The resulting array contained 16 sites. Two sets of hybridizations were performed. First, physically linked targets complementary to probe 1 and probe 2 were used to investigate the cooperative nature of the hybridization signals from each individual target hybridized separately is expected to be similar to the signal from the two unlinked targets hybridized together.

Figure 2. Cooperative hybridization distinguishes between physically linked and unlinked target sequences. Two different probe sequences (blue and pink) are hybridized at the same address on the oligonucleotide array, as described in Materials and Methods. Complementary target sequences (green and purple) are hybridized separately, together but unlabeled or linked in a single molecule. (a) Individual targets hybridize less strongly than linked targets, which hybridize cooperatively. Therefore, the sum of the hybridization signals from two individual targets is less than the signal from the linked targets. (b) The sum of the hybridization signals from each individual target hybridized separately is expected to be similar to the signal from the two unlinked targets hybridized together.

RESULTS

Cooperative hybridization and linkage detection

to determine whether cooperative hybridization could be used to distinguish between linked and unlinked pairs of otherwise identical sequences, we designed a simple experiment to assay for linkage between a pair of 9mer sequences separated by 8 nt. The principle is illustrated in Figure 2.

A 4 × 4 array was synthesized, in which each 400 × 400 µm site contained a mixture of two different 9mer probe sequences, probe 1 (3′-gtgcN1ctcg-5′) and probe 2 (3′-gtanN2ccat-5′). In order to demonstrate that any cooperative effect was sequence-specific, we designed the array to include four variants of each of the probes, in which the central base of probe 1 and probe 2 was patterned onto these sites using MeNPOC photochemistry; the linker is coupled onto these sites. The second set of probes is then patterned onto these sites using MeNPOC photochemistry; the result is an array with a mixture of different DNA probes at every location.

Labeled DNA targets

DNA oligonucleotides bearing a 5′-terminal fluorescein label were synthesized on solid supports using standard phosphoramidite chemistry. Oligonucleotides 10c–27c, 10g–27t, 10c–27t and 10g–27c are based on the sequence 5′Fcc act cac g Ng age tct cca tgc att Ng tattttgct cgg gga gatg cac ggc atga gca-3′, where F denotes fluorescein and N indicates positions 10 and 27. The bases at these positions are indicated in the name of each oligonucleotide. Likewise, oligonucleotides 10c and 10g are based on the sequence 5′-Fct cac g Ng age tct c-3′ and 27c and 27t are based on 5′-Ftgc att gNg tat ttt-3′. The 10c, 10g, 27c and 27t sequences were derived from the position 10 and 27 double variants listed above. In addition to the oligonucleotide targets, human mitochondrial DNAs of 160 bases and 2.5 kb were prepared using single-stranded asymmetric PCR. These DNAs were amplified from samples previously sequenced on an ABI 373A DNA Sequencer. Labeling was by incorporation of biotin-16-dUTP during PCR. Two 2.5 kb amplicons were prepared, differing at three positions. Amplicon 1 had the sequence 93c–1438c–2131a; amplicon 2 had the sequence 93t–1438t–2131g.

**Figure 1.** Synthesis of paired probe arrays. Firstly, a 1:1 mixture of photoprotected and DMT-protected linkers is created over the entire surface of the array. This is accomplished by irradiating a MeNPOC linker photoprotected glass surface with 365 nm UV light, such that half the sites are deprotected (T1/2, half-life of the MeNPOC group). The deprotected sites are reacted with a 5′-DMT-protected nucleoside. Next, the first probe sequence is synthesized at each location on the array: the remaining MeNPOC sites are deprotected and standard light-directed oligonucleotide synthesis is carried out at these sites. When the first probe synthesis has been completed, the 5′-ends are capped. Finally, the second probe sequence is synthesized. Acid deprotection is used to make available the sites previously reserved by DMT protection. After the addition of a photoprotected linker, standard light-directed oligonucleotide synthesis is then used to prepare the second probe sequence.

**Figure 2.** Cooperative hybridization distinguishes between physically linked and unlinked target sequences. Two different probe sequences (blue and pink) are hybridized at the same address on the oligonucleotide array, as described in Materials and Methods. Complementary target sequences (green and purple) are hybridized separately, together but unlabeled or linked in a single molecule. (a) Individual targets hybridize less strongly than linked targets, which hybridize cooperatively. Therefore, the sum of the hybridization signals from two individual targets is less than the signal from the linked targets. (b) The sum of the hybridization signals from each individual target hybridized separately is expected to be similar to the signal from the two unlinked targets hybridized together.

**RESULTS**

Cooperative hybridization and linkage detection

to determine whether cooperative hybridization could be used to distinguish between linked and unlinked pairs of otherwise identical sequences, we designed a simple experiment to assay for linkage between a pair of 9mer sequences separated by 8 nt. The principle is illustrated in Figure 2.

A 4 × 4 array was synthesized, in which each 400 × 400 µm site contained a mixture of two different 9mer probe sequences, probe 1 (3′-gtgcN1ctcg-5′) and probe 2 (3′-gtanN2ccat-5′). In order to demonstrate that any cooperative effect was sequence-specific, we designed the array to include four variants of each of the probes, in which the central base of probe 1 and probe 2 was patterned onto these sites using MeNPOC photochemistry; the linker is coupled onto these sites. The second set of probes is then patterned onto these sites using MeNPOC photochemistry; the result is an array with a mixture of different DNA probes at every location.

Labeled DNA targets

DNA oligonucleotides bearing a 5′-terminal fluorescein label were synthesized on solid supports using standard phosphoramidite chemistry. Oligonucleotides 10c–27c, 10g–27t, 10c–27t and 10g–27c are based on the sequence 5′Fcc act cac g Ng age tct cca tgc att Ng tattttgct cgg gga gatg cac ggc atga gca-3′, where F denotes fluorescein and N indicates positions 10 and 27. The bases at these positions are indicated in the name of each oligonucleotide. Likewise, oligonucleotides 10c and 10g are based on the sequence 5′-Fct cac g Ng age tct c-3′ and 27c and 27t are based on 5′-Ftgc att gNg tat ttt-3′. The 10c, 10g, 27c and 27t sequences were derived from the position 10 and 27 double variants listed above. In addition to the oligonucleotide targets, human mitochondrial DNAs of 160 bases and 2.5 kb were prepared using single-stranded asymmetric PCR. These DNAs were amplified from samples previously sequenced on an ABI 373A DNA Sequencer. Labeling was by incorporation of biotin-16-dUTP during PCR. Two 2.5 kb amplicons were prepared, differing at three positions. Amplicon 1 had the sequence 93c–1438c–2131a; amplicon 2 had the sequence 93t–1438t–2131g.
Figure 3. Design and layout of a paired probe array. (a) Probe 1 was synthesized in four 400 × 1600 µm rectangles, with the central position \( N_1 = A, C, G \) and \( T \) in the successive rectangles. (b) Probe 2 was synthesized in four 400 × 1600 µm rectangles perpendicular to the probe 1 rectangles on the same region of the substrate. The probe 2 central position \( N_2 = A, C, G \) and \( T \) in successive rectangles. (c) The resulting array contained 16 sites, each with a different combination of \( N_1 \) and \( N_2 \) in the two probes.

Figure 4. Fluorescence images of cooperative versus non-cooperative hybridization to paired probe arrays. The design of the array is shown in Figure 3. Unambiguous hybridization to the double perfect match probe pair is shown for four different linked sequence pairs (10g–27c, 10c–27t, 10c–27g and 10g–27t from top of left hand column). Hybridization images of the corresponding unlinked targets are shown in the adjacent right hand column. Because the fluorescence intensity of linked and unlinked target hybridizations varied by as much as two orders of magnitude, images are displayed on two different gray scales, indicated by legends within the figure.

Table 1. Hybridization experiments

<table>
<thead>
<tr>
<th>Array</th>
<th>Target</th>
<th>Conc.</th>
<th>Buffer</th>
<th>Temp</th>
<th>Time</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-mer PPA</td>
<td>oligo 10a–27b</td>
<td>10</td>
<td>A</td>
<td>RT</td>
<td>0.5</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>oligo 10b–27b</td>
<td>10</td>
<td>A</td>
<td>RT</td>
<td>0.5</td>
<td>F</td>
</tr>
<tr>
<td>30-mer PPA</td>
<td>2.5 kb ampiclon 1, 2</td>
<td>0.75</td>
<td>B</td>
<td>37</td>
<td>48</td>
<td>P</td>
</tr>
<tr>
<td></td>
<td>2.5 kb ampiclon 1, 2</td>
<td>0.75</td>
<td>B</td>
<td>37</td>
<td>48</td>
<td>P</td>
</tr>
<tr>
<td>2544 nt 18’</td>
<td>100 mTm Tris</td>
<td>0.5</td>
<td>A</td>
<td>37</td>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>2544 nt 18’</td>
<td>2.5 kb ampiclon 1</td>
<td>0.75</td>
<td>C</td>
<td>40</td>
<td>0h</td>
<td>P</td>
</tr>
</tbody>
</table>

PPA, paired probe array; F: fluorescein; P: phycoerythrin–streptavidin. Buffer A, 60 mM SSPE, 0.005% Triton X-100; buffer B, 2.4 M tetraethylammonium bromide, 10 mM Tris, pH 7.8, 1 mM EDTA, 0.05% Triton X-100; buffer C, 2.4 M methyltriethylammonium bromide, 10 mM Tris, pH 7.8, 1 mM EDTA, 0.05% Triton X-100. Oligo targets a–d are placeholders for different sequences. Actual sequences are given in Figure 4.

Assignment of linkage in a heterozygous mixture

In order to determine if hybridization to paired probe arrays could be used to assign linkage directly in complex heterozygotes, two further experiments were conducted. In each case, equimolar mixtures of two linked targets were hybridized to a 9mer paired probe array. In the first experiment, the mixture consisted of 10c–27t and 10g–27c. In the second experiment, the mixture was of 10g–27t and 10c–27c. Although the two experiments have targets that are identical in sequence composition, the pairing is different. This is clearly detected in the experiment, which allows the pairings (linkages) to be determined in each case. The panel in the right hand column shows a hybridization image of 10c, 10g, 27c and 27t. The sequence composition is identical to the two panels in the left hand column. However, in this case the individual targets are unlinked and hence no cooperative effect is observed. Because the fluorescence intensity of linked and unlinked target hybridizations varied by as much as two orders of magnitude, images are displayed on two different gray scales, indicated by legends within the figure.
hybridization, in which unlinked targets have the same sequence composition as the linked targets, shows lower signal and no evidence of cooperativity (Fig. 5, right hand panel). These results show that paired probe arrays can be used to assign linkage in mixtures containing two different multiply polymorphic alleles.

**Cooperative hybridization over distances of >2 kb and construction of SNP haplotypes**

To examine the strength and specificity of linkage over a greater distance, we synthesized paired probe arrays with the probe sequences chosen from different regions of a 2.5 kb mitochondrial DNA amplicon. The length of the probes was increased to 30mers to allow hybridizations to be performed under more stringent conditions. Higher stringency was used to reduce the secondary structure in the targets and to favor cooperative hybridization by destabilizing individual hybridizations.

Arrays of three different designs were synthesized. In each design a different pair of single nucleotide polymorphisms (SNPs) was interrogated. In each of three experiments shown, a 50:50 mixture of two 2.5 kb target amplicons was analyzed on a paired array (Fig. 6). The two 2.5 kb amplicons are from the
identical region of human mitochondrial DNA, but are polymorphic and differ from each other at the specific sites analyzed by the arrays. In the first experiment, SNPs at positions 1438 and 2131, separated by 693 nt, were analyzed. The second experiment queried SNPs 1345 nt apart, at positions 93 and 1438. The third experiment queried SNPs 2038 nt apart, at positions 93 and 2131. In each experiment, the correct probe cells have the highest intensity (Fig. 6). The results show single base mismatch discrimination and linkage detection between loci separated by distances up to 2 kb. However, the discrimination is not as good as in the 9mer paired probe array. This is not surprising, as the discrimination is based on differences of two bases out of 60, as opposed to two bases out of 18 in the 9mer experiments.

The data were then analyzed by comparing each probe pair with its six single base alterations (the three single base changes in probe 2, keeping probe 1 constant and the three single base changes in probe 1, keeping probe 2 constant) (Fig. 6, right hand column). In this analysis, the distinction between the correct linkage assignments and the incorrect ones is even more apparent. A score of 1 on the discrimination chart means the intensity at that position on the array was the same as the average intensity of all its one base alterations, i.e. the discrimination score for the probe pairing a–a = (a–a)/[(a–c + a–g + a–t + c–a + g–a + t–a)/6]. All of the incorrect linkage assignments are close to or <1, while the correct linkage assignments produce values of ≥1.9. As with the 9mer paired probe array, the correct linkage assignments (93c–1438c, 93t–1438t, 1438c–2131a, 1438t–2131g, 93c–2131a and 93t–2131g) were easily distinguished from the incorrect phase (93c–1438t, etc.). Finally, by combining the data from the pairwise experiments, the two haplotypes 93c–1438c–2131a and 93t–1438t–2131g can be unambiguously reconstructed.

**Sequence independence of cooperative effect**

The experiments described above were carried out using a small number of specific pairs of sequences. We designed two more arrays to investigate cooperative hybridization with a much larger number and variety of different sequences. The aim was to determine to what extent the cooperative effect we observed was sequence-dependent, in order to assess if it could be extrapolated to a larger set of probes.

In the first experimental design, six different 12mer sequences were each paired with a resequencing array containing 86 overlapping 15mer probes. Therefore, each of these resequencing arrays contained a constant 12mer sequence to act as an ‘anchor’ for the target, which was queried by the variable 15mer probes in the array. Each 15mer was represented by four single base substitutions (A, C, G and T substitutions at the central position) to give a total of 2064 combinations (6 anchors × 86 probes × 4 substitutions). Of these, 516 were perfect matches to both probes in the pair and the remainder contained a single base mismatch at the central position in the 15mer probe.

A 160 nt fluorescein-labeled amplicon was hybridized to the array and an image of the hybridization pattern was obtained (Fig. 7). The array included an unpaired tiling as a control. Every paired tiling gave a greater signal than the control tiling. The extent of this improvement, in overall intensity and discrimination, is related to the hybridization strength of the region represented by the constant 12mer anchor sequence. For example, positions 1–12 have very low intensity in the control tiling and anchor 1–12 has a small effect on the intensity and specificity of the hybridization in its paired tiling. Positions 15–26 perform well in the control and have a dramatic impact as an anchor sequence. No signal increase was expected in the regions where interrogation probe sequences overlap with the anchor sequence (outlined in Fig. 7), because each target molecule cannot hybridize to more than one probe in these cells.

The results showed that a variety of different anchors can be paired with the same resequencing tiling with similar effect, that a given anchor sequence can pair cooperatively with a variety of different sequences at various distances from the anchor and that signal enhancement by the anchor probe is related to the hybridization strength of the anchor sequence.
In order to extend these observations and to examine probe–anchor interactions over greater distances, we next paired a single 20mer anchor probe with a 20mer tiling array interrogating 2544 nt of the human mitochondrial genome. Two separate arrays were synthesized. The first contained only the 10 176 20mers (2544 × 4 substitutions per position) as a control. In the second array each cell was a mixture of one of the 10 176 probes and the 20mer complementary to positions 1427–1446 on the amplicon.

An analysis of signal intensities versus position in the 2544 base sequence reveals the characteristic dip in the region where the anchor overlaps with the variable probe and increased signal intensity and discrimination elsewhere on the array (an average of 15 times fluorescence intensity over the unpaired control) (Fig. 8). At separations >1000 bases between the anchor and probe sites on the target, a strong cooperative binding effect can still be seen. Under the conditions used, 97% of the sequence (2459 of 2544 bases) could be determined by simply identifying the probe with the greatest intensity in each set of A, C, G and T substitution probes. In contrast, under these stringent conditions only 94% of the sequence (2128 bases) could be determined on the unpaired control array.

**DISCUSSION**

In this paper we have begun to explore a new approach for analyzing nucleic acids by hybridization to oligonucleotide arrays. By synthesizing two probe sequences at the same address on the array, it is possible to hybridize to two target sequences simultaneously. We showed that physically linked targets hybridize cooperatively, resulting in an increase in binding strength. This property was used to detect physical linkage between pairs of target sequences and to define haplotypes in 50:50 mixtures of polymorphic sequences.

Typically, linkage analysis is carried out over relatively large distances (e.g. megabases to hundreds of kilobases). However, maps are being made at increasingly higher resolution and SNP variation in the human genome occurs at the rate of ~1 SNP/kb. Also, different kinds of mapping are carried out at many resolutions depending on the need. For example, sequencing often requires special efforts to determine sequence order over a few kilobases or less. This is useful in completing sequence assembly when local regions of sequence are undefined (e.g. due to cloning difficulties).

We have shown that cooperative hybridization can be used to detect interactions over a distance of 2 kb with single base discrimination. Importantly, we demonstrated the reconstruction of haplotypes based on the information obtained from pairwise linkage experiments. From these results, it is clear that the approach has the resolution needed to construct physical maps over distances of at least several kilobases. As the distance between regions of interest on the target increases, hybridization conditions needed to be made more stringent. Various tetraalkyl ammonium salts were used to increase stringency while maintaining a hybridization temperature between 20 and 40°C (11). This allowed the single probe hybridizations to be selectively destabilized, increasing the discrimination between dual and single probe hybrids.

The potential to extend these observations to much greater distances exists, but there are some technical obstacles to be overcome. One difficulty is in amplifying and obtaining long ssDNA targets. In addition, detecting interactions over much greater distances will likely require higher affinity probes. This can be achieved by increasing the length of the individual probes. Alternatively, three or more probes could be synthesized at each site. By making use of enzymatic methods of mismatch discrimination (12,13), it may be possible to retain single base mismatch discrimination while achieving much higher hybrid affinities.

An important question was to what extent the cooperative binding to arrayed probes was sequence-dependent. By combining an ‘anchor’ probe with different resequencing tiling arrays, we showed that most probe sequences can be used to examine linkage in a paired array. Six different 12mer probes performed similarly with 86 different probe sets and one 20mer probe was shown to enhance hybridization when paired with -2000 different 20mers. While the ability of a probe pairing to hybridize cooperatively was related to the probe’s performance as a single probe, each probe’s hybridization yield and discrimination were
enhanced by being part of a paired array. Furthermore, the method is general in that any combination of probes can be made at each and every site on the array. While the synthesis of all probes proceeds in parallel, the synthesis of any individual probe is addressed independently of the others.

Further applications of this method include increasing the specificity of probe–target interactions, in order to enhance the discrimination between specific and non-specific hybridizations. This could be used to improve the sensitivity and specificity of resequencing and gene expression monitoring from complex mixtures of target sequence (e.g. total mRNA populations and total genomic DNA) (14,15). The technique may also have potential for measuring short distances and could possibly be used for comparative counting or sizing of dinucleotide or other repetitive sequences. A potentially interesting application is sequence reconstruction (contig assembly) in sequencing by hybridization (SBH). One of the major obstacles to using short DNA probes for SBH is the presence of repeated sequences. De novo SBH utilizes an array of the set of all probes of length n (usually 6–10) to determine the presence of length n subsequences in the target of interest. These subsequences are compared with each other and aligned to reconstitute the contiguous target sequence. Repeated regions of length n or longer complicate the sequence reassembly, since it is no longer possible to unambiguously determine the linkage of sequences on either side of the repeat. It has been proposed that measuring the distance between reference points in a sequence could increase the possible fragment length an array can resequence by >4-fold (16).

ACKNOWLEDGEMENTS

We would like to thank Don Morris for valuable insights into array-based hybridization behavior, Jonathan Forman, Mark Trulson and David Lockhart for advice and discussion and Jo Ann Iwasa and Kevin Johnson for help with chip synthesis. We also thank Teena Edwards, Andy Lau and Fay Caldwell for help with the figures and manuscript preparation. This work was in part supported by NIH grant 5P01 HG01 323.

REFERENCES